モンゴル語全文検索システムの実現

満　都拉[†,††]　　藤井　敦[†,†††]　石川　徹也[†]
[†]筑波大学
[††]図書館情報大学
[†††]科学技術振興機構 CREST
{mandula,fujii,ishikawa}@slis.tsukuba.ac.jp

1．はじめに

今日の高度情報化時代においても、伝統的な縦書きモンゴル語による電子化されたデータベースは未だ存在しない。最大の原因は、モンゴル語の文字変形が複雑でコンピュータ上で扱う方式が確立していないためである。

筆者らは、モンゴル語をコンピュータ上で扱う方式を確定するため、モンゴル語文字コードの問題を解決し、モンゴル語入出力インタフェースを実現した[10]。

本研究では、この入出力インタフェースを利用してモンゴル語新聞記事を電子化し、全文検索システムを構築した。

以下、２.で全文検索システムについて説明し、３.で既存のモンゴル語文字コードと本研究の文字コードについて説明し、４.で入出力インタフェースについて説明する。

２．モンゴル語全文検索システム

全文検索システムは、文書集合全体から検索質問に適合する文書を探し出すシステムである。

本研究で実現したモンゴル語全文データ検索システムの構成を図１に示す。

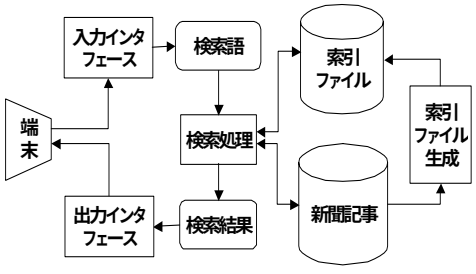


図1　モンゴル語全文検索システムの構成

本研究で利用するモンゴル語全文データを作成するために、筆者らが提案した入力方式[10]によって、内モンゴル自治区においてモンゴル語で発行されている「内蒙古日報」の新聞記事を３０件を人手で入力した。

全文検索システムの検索方法には逐次検索と索引検索の２種類がある[8]。逐次検索は検索質問を与える度に検索質問と文書内の文字を１文字ずつ照合することによって検索質問に適合する文書を探し出す。このため、検索する対象データが大規模化すると効率が悪い。

索引検索は、検索する対象データに予め索引を作成し、索引を用いて検索を行うため、対象データの規模が大きくなっても効率が低下しない。本研究では検索システムの効率を重視して索引検索方法を採用した。

2.1 索引ファイルの生成

図２や図４に示すように、モンゴル語は分かち書きされているため、語の検出は日本語より容易である。よって、文書の中の単語を特定して抽出することにより索引を生成できる。しかし、文書の中から抽出された単語の全てが文書を特徴付ける索引とは限らない。特に、語と語の関係を表す機能語は文書の内容を特徴付けることができない。例えば、助詞や助動詞などが単語として抽出されても索引語には相応しくないので、不要語と呼ぶ。文書の内容を特徴つける上で重要な内容を表すことができる単語を内容語と呼び、索引語に相応しい。この索引語を抽出する作業を索引ファイルの生成と言う。

モンゴル語は表音文字であり、単語が英語のように空白で区切られている。この空白によって単語を抽出することができる。しかし、モンゴル語の助詞、助動詞などの文書の内容を特徴付ける上で重要ではない語を不要語として削除しなければ検索効率に影響する。そこで、本研究では不要語リスト（表１）を作成した。

表１　不要語リスト

モンゴル文字	日本語の意味	モンゴル文字	日本語の意味
	の		を
	に		の
	で		の
	で		でした
	を		です
	ついて		なる
	と		なった
	だった		と
	だ		と
	を		と

2.2 検索質問の処理

実際の検索エンジンを調査した結果、入力される検索キーワードの数が少ない。例えば、インタネット上の WWW サイトの検索エンジン Excite（http://www.excite.com）に入力される索引語の平均は 2.35 である[4]。よって、本システムでも、ユーザが少数のキーワードを入力することを想定している。

ユーザがキーワードを入力する時に誤りがないとは限らないため、入力したキーワードが正しいかどうかを確認する必要がある。ユーザがアルファベット（モンゴル語の読み）を入力しリターンキーによって確定するごとに、システムは入力されたアルファベットをモンゴル文字に変換する。この際、アルファベットとモンゴル文字をテキストボックス上に併記して表示する。この結果、ユーザは画面上のモンゴル文字とアルファベットの綴りが合っているかどうかを確認することができる。

2.3 検索モデル

検索モデルには、ブーリアンモデルを採用した。しかし、ブーリアンモデルでは検索結果に順次付けることが出来ない。文書の中で出現頻度が高い語は、その文書の内容に対して重要である。そこで、本システムでは索引ファイルを生成するとき、索引語の出現頻度を予め索引ファイルに付与する。よって、出現頻度の高い方から出力する。

2.4 検索結果

検索結果は検索質問に適合する新聞記事を ID、新聞名、年月日、曜日、タイトル、記事の順に表示する。例えば、図２のようになる。

図2　検索結果の例

このシステムを実現するためには、先ずモンゴル語文字コードの問題を解決し、入出力インタフェースを実装しなければならない。そこで、以下３．で文字コードの問題を解決する方法について説明し、４．で入出力インタフェースの実装方法について説明する。

３．モンゴル語文字コード

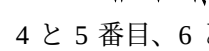
自然言語によるデータを電子化して情報処理を行うとき、先ずその言語を表記する記号である文字コードの設定が必要である。既存の文字コードを分析した結果、文字コードに様々な欠点があり、モンゴル語の全ての情報を電子的に保存できないことがわかった。そこで、既存の文字コードを利用せず、モンゴル語の読みをローマ字で入力してアスキーコードで保存する方式を提案した[10]。その結果、モンゴル語の発音情報を失わず保存でき、情報処理に適したデータベースが実現できる。

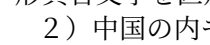
4．モンゴル語入出力インタフェース

データベースの構築には、入出力インタフェースが重要である。入出力インタフェースによってキーボードからデータを入力したり、コンピュータ中のデータを出力できることが重要である。この入出力インタフェースの実装において、重要なのは入力方式である。

4.1 既存の入力方式

代表的な入力方式として、次の 2 種類を紹介する。

1）先ずドイツのベルリン自由大学の Corff Oliver が提案した入力方式（Corff 方式）がある[7]。この方式ではモンゴル語の７つの母音を５つの文字で入力して、形が同じ文字を一つの文字として扱う。モンゴル語では 7 つの母音  において 4 と 5 番目、6 と 7 番目の形が同じでもそれぞれ発音が異なる。そのため Corff 方式は同形異音文字を区別できないという問題がある。

2）中国の内モンゴル大学で作られた入力方式（内大方式）がある[1]。内大方式ではモンゴル語の 7 つの母音  を区別して A, E, I, 0, O, V, U で入力する。また、子音をそれぞれ区別して入力するので同形異音文字を区別することができる。しかし、この方式では同形異音文字を区別するために数字のゼロを用いて、4 番目の母音  を入力しているため、文字と数字が混在するデータを扱うことが困難である。

4.2 本研究の入力方式

4.1 で説明したように既存の入力方式を直接利用すると多少問題がある。しかし、内大方式は母音と子音の同形異音文字を区別しており、モンゴル語の特徴を表現することができる。よって、この入力方式を改良することで利用価値が高まる。そこで、本研究では内大方式を改良して利用することにした。内大方式の  をアルファベット大文字の O で入力する。キーボード入力するとき利便性のため、内大方式のほとんどがアルファベットの大文字で入力しているのを、本研究では全て小文字で入力する。ただし、大文字はモンゴル文字の特殊な形を表すために用いる。例えば、ｄとｔの語中形  や語尾形の をそれぞれＤとＴで入力する。また、文字の表記をなるべくアルファベットに限定して他の記号を用いないようにするために、子音 はｘで入力する。内大方式ではこの文字を$で入力する。

モンゴル語は表音文字であり、また文字が語頭、語中、語尾の位置によって形が変わるため、入力の時に語中の位置を指定する必要がある。本研究では、スペースキーによって単語を区切る。即ち、スペースの後ろにある文字は語頭形で、スペースの前に入る文字は語尾形を表して、前後に文字がつながっている文字は語中形と判断することができる。

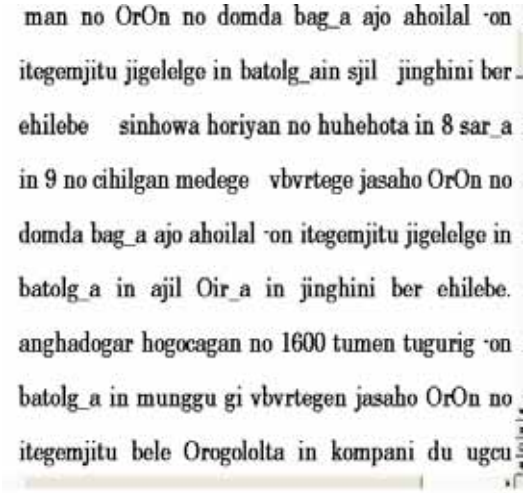

man no OrOn no domda bag_a ajo ahoilal -on
itegemjitu jigelelge in batolg_ain sjil jinghini ber
ehilebe sinhowa horiyan no huhehota in 8 sar_a
in 9 no cihilgan medege vbvrtege jasaho OrOn no
domda bag_a ajo ahoilal -on itegemjitu jigelelge in
batolg_a in ajil Oir_a in jinghini ber ehilebe.
anghadogar hogocagan no 1600 tumen tugurig -on
batolg_a in munggu gi vbvrtegen jasaho OrOn no
itegemjitu bele Orogololta in kompani du ugcu

図 3　データ保存形式の例

そこで、改良した方式によって入力されたデータをモンゴル語で出力する。しかし、保存する方式に既存の文字コードを利用すると同形異音文字を区別できない。特に、アリガリ文字が表現できないなどの問題がある。そこで、本研究では同形異音文字を区別し、アリガリ文字も扱うため、モンゴル語の読みをローマ字で入力する方式とした。この結果、モンゴル語の読み情報が失われず、同形異音文字を区別でき、アリガリ文字の情報も保存できる。

この方式による入力データの内容は図３のようにアスキー文字で保存される。出力画面

は図 4 のようになる。

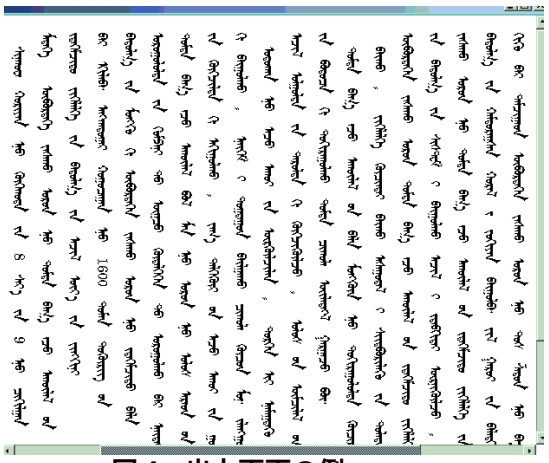


図 4　出力画面の例

5.終わりに

本研究では、モンゴル語の読みによる入出力インタフェースを実装し、モンゴル語全文検索システムを実現した。しかし、本システムではモンゴル語の発音情報に注目して検索するため、字形による検索を行うことができない。

今後の研究課題として、データ量を増やして、検索システムの性能評価を行う必要がある。また、発音情報を扱うだけではなく字形による検索も可能にする必要がある。

参考文献

[1] Ochir ら.WINDOWS 環境でのモンゴル文字入力方法の検討，内蒙古大学学報，pp. 102—108，2000.

[2] 上村明.モンゴル語処理の現状，bit. Vol. 30，No. 6，pp.70—71，1998.

[3] 亀井孝、河野六郎、千野栄一．言語学大辞典 第 4 巻 世界言語編（下—2），三省堂，1992.

[4] 北研二、津田和彦、獅獅堀正幹． 情報検索アルゴリズム， 共立出版株式会社，2002.

[5] 確精扎布.蒙古文編碼，内蒙古大学出版社，2000.

[6] 徳永健伸， 情報検索と言語処理， 東京大学出版会， 1999.

[7] 中里致元． モンゴル語電子化計画，http://texa.human.is.tohoku.ac.jp/~chigen/2m_dsv_j.htm.

[8] 馬場肇． 日本語全文検索システムの構築と活用， ソフトバンク株式会社出版事業部， 1998.

[9] 満都拉．モンゴル語専門用語の由来分析，第 13 回専門用語研究シンポジウム，pp. 13—20，2000.

[10] 満都拉、藤井敦、石川徹也． モンゴル語入出力インタフェースの実現と書誌データ検索への応用， 言語処理学会第 8 回年次大会発表論文集， pp.184 — 187，2002.